2004 Proceedings of the Autumn Conference, Korean Statistical Society

Partitioning likelihood method in the analysis of non-monotone missing data

Jae-Kwang Kim

ABSTRACT

We address the problem of parameter estimation in multivariate distributions under ignorable non-monotone missing data. The factoring likelihood method for monotone missing data, termed by Rubin (1974), is extended to a more general case of non-monotone missing data. The proposed method is algebraically equivalent to the Newton-Raphson method for the observed likelihood, but avoids the burden of computing the first and the second partial derivatives of the observed likelihood. Instead, the maximum likelihood estimates and their information matrices for each partition of the data set are computed separately and combined naturally using the generalized least squares method. A numerical example is also presented to illustrate the method.

KEY WORDS. *EM algorithm, Gauss-Newton method, Generalized least squares, Missing at random.*

1. Introduction

Let (Y_{1i}, Y_{2i}) be a vector of bivariate normal random variable distributed as

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \overset{iid}{\sim} N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \right], \tag{1}$$

where $\overset{iid}{\sim}$ is the abbreviation of independently and identically distributed. We assume that the observations are missing at random (MAR) in the sense of Rubin (1976) so that the relevant likelihood is the observed likelihood, the marginal likelihood of the observed data.

To estimate the parameters, a direct maximum likelihood method that maximizes the observed likelihood can be used. To do this, we need to compute the observed likelihood and its partial derivatives. The Newton-Raphson type solution to the likelihood equation also requires the computation of the second-order partial derivatives. The computation of the first and the second partial derivatives can be cumbersome.

[1]Department of Applied Statistics, Yonsei University, Seoul, 120-749.

The factoring likelihood method, termed by Rubin (1974), avoids the burden of computing the partial derivatives and still compute the maximum likelihood estimate (MLE) of the observed likelihood. The factoring likelihood method computes MLEs easily, but is applicable only to the monotone missing data. For the definition of monotone missing pattern and the non-monotone missing pattern, see Little and Rubin (2002, section 1.2). Under monotone missing pattern, the observed likelihood can be factored into the marginal likelihood and the conditional likelihood so that the maximum likelihood estimates can be estimated separately at each likelihood. For example, assume that Y_1 is fully observed with n observation and Y_2 is subject to missing with $r(< n)$ observation. Anderson (1957) first consider the estimation of parameters under this setup by using an alternative representation of the bivariate normal distribution as

$$\begin{aligned} Y_{1i} &\overset{iid}{\sim} N(\mu_1, \sigma_{11}) \\ Y_{2i} \mid (Y_{1i} = y_{1i}) &\overset{iid}{\sim} N(\beta_{20\cdot1} + \beta_{21\cdot1} y_{1i}, \sigma_{22\cdot1}), \end{aligned} \tag{2}$$

where $\beta_{20\cdot1} = \mu_2 - \beta_{21\cdot1}\mu_1$, $\beta_{21\cdot1} = \sigma_{11}^{-1}\sigma_{12}$ and $\sigma_{22\cdot1} = \sigma_{22} - \beta_{21\cdot1}^2\sigma_{11}$. The observed likelihood is then written as a product of marginal likelihood of a fully observed variable Y_1 and the conditional likelihood of Y_2 given Y_1. Thus, the parameters μ_1 and σ_{11} for the marginal distribution of Y_1 can be estimated with n observation and the other regression parameters, $\beta_{20\cdot1}$, $\beta_{21\cdot1}$, and $\sigma_{22\cdot1}$, can be estimated from the conditional distribution with r observation.

Note that the factoring likelihood approach consists of two steps. In the first step, the likelihood is factored, and in the second the MLE for each likelihood is computed separately. The advantage of the factoring likelihood method is that the MLEs are easily computed because the marginal and the conditional likelihoods are of known form and thus we can directly use the known solutions of the likelihood equations for each likelihood. Rubin (1974) recommended the factoring likelihood approach as a general framework in the analysis of missing data with monotone missing pattern.

Under the non-monotone missing pattern, the factoring likelihood approach is not directly applicable because the parameters are no longer orthogonal. The EM algorithm, proposed by Dempster, Laird and Rubin (1977), can be used to compute the MLEs under the general missing pattern, but uses an iterative procedure and does not provide the information matrix directly.

In this paper, we consider an extension of the factoring likelihood method, called the partitioning likelihood method, to the analysis of non-monotone missing data. Note that the

factoring likelihood method ease the computation of the MLEs but the resulting estimators are no longer independent because of the non-orthogonality of the parameters. Thus, in addition of the two steps in the original factoring likelihood approach, we need another step to combine these separate MLEs computed within each likelihood to produce the final MLEs. The proposed method turns out to be essentially the same as the direct maximum likelihood method using the Newton-Raphson algorithm but has some computational advantages. The proposed method is described under the bivariate normal setup in Section 2. A justification of the proposed method under a more general setup is made in Section 3. The proposed method is applied to a categorical data example in Section 4.

2. One-step estimator

The proposed method can be described into three steps:

[Step 1] Partition the original sample into several disjoint sets according to the missing pattern.

[Step 2] Compute MLE for each identified parameters separately in each partition of the sample.

[Step 3] Combine the estimators to get a set of final estimates using a generalized least squares (GLS) form.

In Step 1, with a non-monotone missing pattern with two variable, we have three types of respondents that contain information about the parameters. The first set H of units have both Y_1 and Y_2 observed, the second set K of units have Y_1 observed but are missing Y_2, and the third set L of units have Y_2 observed but are missing Y_1. That is, we partition the sample into several disjoint sets according to the pattern of missingness. We also define M to be the set of units that have both Y_1 and Y_2 missing. Let n_H, n_K, n_L, and n_M be the sample size of the set H, K, L, and M, respectively. Note that $n = n_H + n_K + n_L + n_M$.

In Step 2, we obtain the following estimators in each set: For set H, we get the ML estimates for the five parameters in (2): $\hat{\beta}_{20\cdot1,H}$, $\hat{\beta}_{21\cdot1,H}$, $\hat{\sigma}_{22\cdot1,H}$, $\hat{\mu}_{1,H}$, and $\hat{\sigma}_{11,H}$. For the set K, the ML estimates $\hat{\mu}_{1,K}$ and $\hat{\sigma}_{11,K}$ are obtained for μ_1 and σ_{11}, respectively. For the set L, the ML estimates $\hat{\mu}_{2,L}$ and $\hat{\sigma}_{22,L}$ are obtained for $\mu_2 = \beta_{20\cdot1} + \beta_{21\cdot1}\mu_1$ and $\sigma_{22} = \sigma_{22\cdot1} + \beta_{21\cdot1}^2\sigma_{11}$, respectively.

In Step 3, we use the GLS method to combine the nine estimators into an estimator for

the five parameters. The nine estimates are

$$\hat{\boldsymbol{\eta}} = \left(\hat{\beta}_{20\cdot1,H}, \hat{\beta}_{21\cdot1,H}, \hat{\sigma}_{22\cdot1,H}, \hat{\mu}_{1,H}, \hat{\sigma}_{11,H}, \hat{\mu}_{1,K}, \hat{\sigma}_{11,K}, \hat{\mu}_{2,L}, \hat{\sigma}_{22,L}\right)'. \tag{3}$$

The expected values of the nine estimates are

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \left(\beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1}, \mu_1, \sigma_{11}, \mu_1, \sigma_{11}, \beta_{20\cdot1} + \beta_{21\cdot1}\mu_1, \sigma_{22\cdot1} + \beta_{21\cdot1}^2\sigma_{11}\right)' \tag{4}$$

and the asymptotic covariance matrix is

$$\mathbf{V} = diag\left\{\Sigma_{bb}, \frac{2\sigma_{22\cdot1}^2}{n_H}, \frac{\sigma_{11}}{n_H}, \frac{2\sigma_{11}^2}{n_H}, \frac{\sigma_{11}}{n_K}, \frac{2\sigma_{11}^2}{n_K}, \frac{\sigma_{22}}{n_L}, \frac{2\sigma_{22}^2}{n_L}\right\}, \tag{5}$$

where $\boldsymbol{\theta} = (\beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1}, \mu_1, \sigma_{11})$ and

$$\Sigma_{bb} = \begin{pmatrix} n_H^{-1}\sigma_{22\cdot1}\left(1 + \sigma_{11}^{-1}\mu_1^2\right) & -n_H^{-1}\sigma_{11}^{-1}\sigma_{22\cdot1}\mu_1 \\ -n_H^{-1}\sigma_{11}^{-1}\sigma_{22\cdot1}\mu_1 & n_H^{-1}\sigma_{11}^{-1}\sigma_{22\cdot1} \end{pmatrix}.$$

The GLS formulation in (4) and (5) is a nonlinear model of the five parameters. Using a Taylor expansion on the nonlinear model, a step of the Gauss-Newton method can be formulated as

$$\mathbf{e}_\eta = \mathbf{X}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_S\right) + \mathbf{u}, \tag{6}$$

where $\mathbf{e}_\eta = \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\left(\hat{\boldsymbol{\theta}}_S\right)$, $\hat{\boldsymbol{\theta}}_S$ is the initial estimator of $\boldsymbol{\theta}$, $\boldsymbol{\eta}\left(\hat{\boldsymbol{\theta}}_S\right)$ is the vector (4) evaluated at $\hat{\boldsymbol{\theta}}_S$,

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & \mu_1 & 2\beta_{21\cdot1}\sigma_{11} \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & \beta_{21\cdot1} & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & \beta_{21\cdot1}^2 \end{pmatrix}', \tag{7}$$

and, approximately,

$$\mathbf{u} \sim (\mathbf{0}, \mathbf{V}),$$

where $\mathbf{V}$ is the covariance matrix defined in (5). For a brief description of the Gauss-Newton method for the estimation of nonlinear models, see Fuller (1996, section 5.5).

The procedure can be carried out iteratively until convergence, but we used a single step of the procedure. For a suitable choice of the initial estimates, the one-step estimator is a very good approximation to the maximum likelihood estimator. The estimator is

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_S + \left(\mathbf{X}_S'\hat{\mathbf{V}}_S^{-1}\mathbf{X}_S\right)^{-1}\mathbf{X}_S'\hat{\mathbf{V}}_S^{-1}\mathbf{e}_\eta, \tag{8}$$

where $\mathbf{X}_S$ and $\hat{\mathbf{V}}_S$ are evaluated from $\mathbf{X}$ in (7) and $\mathbf{V}$ in (5), respectively, using the initial values of θ. The covariance matrix of the estimator in (8) can be estimated by

$$\mathbf{C} = \left(\mathbf{X}_S' \hat{\mathbf{V}}_S^{-1} \mathbf{X}_S\right)^{-1}. \tag{9}$$

The initial values for the iterative procedure are $\hat{\beta}_{20\cdot 1,H}$, $\hat{\beta}_{21\cdot 1,H}$, $\hat{\sigma}_{11\cdot 2,H}$,

$$\tilde{\mu}_1 = (n_H + n_K)^{-1} (n_H \bar{y}_{1,H} + n_K \bar{y}_{1,K}),$$

and

$$\tilde{\sigma}_{11} = (n_H + n_K - 2)^{-1} \left[(n_H - 1) s_{1,H}^2 + (n_K - 1) s_{1,K}^2\right],$$

where $\bar{y}_{1,H}$ and $\bar{y}_{1,K}$ are the sample means of Y_1 in the sets H and K, respectively, and $s_{1,H}^2$ and $s_{1,K}^2$ are the sample variances of Y_1 in the sets H and K.

3. Justification

Let the score function of a likelihood be defined as

$$S(\mathbf{y};\theta) = \partial \log l(\theta) / \partial\theta,$$

where $l(\theta) = \prod_i f(\mathbf{y};\theta)$ is the likelihood function of parameter θ. The maximum likelihood estimator of θ can be defined as a solution to the Newton-Raphson variant of scoring method

$$\theta^{(k+1)} = \theta^{(k)} + \left[I\left(\theta^{(k)}\right)\right]^{-1} S\left(\mathbf{y};\theta^{(k)}\right), \tag{10}$$

where $I(\theta) = E\left[-\partial^2 \log l(\theta) / \partial\theta^2\right]$ is the expected information matrix for θ. It is known that if the starting value is a $\sqrt{n}$-consistent estimator of θ, then one-step iterate $\theta^{(1)}$ in (10) is asymptotically equivalent to the maximum likelihood estimator of θ. (e.g. Lehmann, 1983, Theorem 3.1, p. 422.)

Now, under the missing data structure in Section 2, we show that the one-step estimator in (8) is equivalent to the Newton-Raphson solution in (10). Note that the observed log-likelihood can be written as a sum of the log-likelihood in each set:

$$\log l(\theta) = \log l_H(\theta) + \log l_K(\theta) + \log l_L(\theta), \tag{11}$$

where $l_H = \prod_{i\in H} f(\mathbf{y};\theta)$ is the likelihood function defined in set H, and l_K and l_L are defined similarly. Under MAR, l_H is the likelihood for the joint distribution of Y_1 and Y_2, l_K is the likelihood for the marginal distribution of Y_1, and l_L is the likelihood for the marginal distribution of Y_2. By (11), the score function for the likelihood can be written as

$$S(\mathbf{y};\theta) = S_H(\mathbf{y};\theta) + S_K(\mathbf{y};\theta) + S_L(\mathbf{y};\theta) \tag{12}$$

and the expected information matrix also satisfies the additive decomposition:

$$I(\theta) = I_H(\theta) + I_K(\theta) + I_L(\theta), \tag{13}$$

where $I_H(\theta) = E\left[-\partial^2 \log l_H(\theta)/\partial\theta^2\right]$, and $I_K(\theta)$ and $I_L(\theta)$ are defined similarly. Let $\eta_H = \eta_H(\theta)$ be a parametrization that $I_H(\eta_H)$ matrix is easy to compute. One such parametrization is $\eta_H = (\eta_{H1}, \eta_{H2})$, where η_{H1} is the parameters for the conditional distribution and η_{H2} is the parameters for the marginal distribution. Since the parameters for the conditional distribution are orthogonal to those for the marginal distribution, the parametrization $\eta_H = (\eta_{H1}, \eta_{H2})$ makes the $I_H(\eta_H)$ matrix block-diagonal. The parametrization for the set H need not be the same as that for the set K nor for the set L, providing more flexibility in choosing the parametrization. Separate orthogonal parametrization in each set will lead to computational advantages over the direct maximum likelihood method.

The equation in (13) can be written as

$$\begin{aligned} I(\theta) &= \left(\frac{\partial\eta_H}{\partial\theta}\right) I_H(\eta_H) \left(\frac{\partial\eta_H}{\partial\theta}\right)' + \left(\frac{\partial\eta_K}{\partial\theta}\right) I_K(\eta_K) \left(\frac{\partial\eta_K}{\partial\theta}\right)' + \left(\frac{\partial\eta_L}{\partial\theta}\right) I_L(\eta_L) \left(\frac{\partial\eta_L}{\partial\theta}\right)' \\ &= X'\hat{V}^{-1}X, \end{aligned} \tag{14}$$

where $X' = \left(\frac{\partial\eta_H}{\partial\theta}, \frac{\partial\eta_K}{\partial\theta}, \frac{\partial\eta_L}{\partial\theta}\right)$ and $\hat{V}^{-1} = diag\{I_H(\eta_H), I_K(\eta_K), I_L(\eta_L)\}$. Now, consider the score function in (12). Using the chain rule, the score function can be written as

$$S(\mathbf{y};\theta) = \left(\frac{\partial\eta_H}{\partial\theta}\right) S_H(\mathbf{y};\eta_H) + \left(\frac{\partial\eta_K}{\partial\theta}\right) S_K(\mathbf{y};\eta_K) + \left(\frac{\partial\eta_L}{\partial\theta}\right) S_L(\mathbf{y};\eta_L). \tag{15}$$

Let $\hat{\eta}_H$ be the MLE of the likelihood l_H. Taking a Taylor expansion of $S_H(\mathbf{y};\eta_H)$ about $\hat{\eta}_H$ leads to

$$S_H(\mathbf{y};\eta_H) \doteq S_H(\mathbf{y};\hat{\eta}_H) - \mathcal{I}_H(\hat{\eta}_H)(\eta_H - \hat{\eta}_H),$$

where $\mathcal{I}_H(\eta_H) = -\partial^2 \log l_H(\eta_H)/\partial\eta_H^2$. Using $S_H(\mathbf{y};\hat{\eta}_H) = 0$ and the weak convergence of the observed information matrix to the expected information matrix, we have

$$S_H(\mathbf{y};\eta_H) \doteq -I_H(\hat{\eta}_H)(\eta_H - \hat{\eta}_H).$$

Similar results hold for the sets K and L. Thus, (15) becomes

$$\begin{aligned} S(\mathbf{y};\theta) &\doteq \left(\frac{\partial\eta_H}{\partial\theta}\right) I_K(\hat{\eta}_H)(\hat{\eta}_H - \eta_H) + \left(\frac{\partial\eta_K}{\partial\theta}\right) I_K(\hat{\eta}_K)(\hat{\eta}_K - \eta_K) \\ &\quad + \left(\frac{\partial\eta_L}{\partial\theta}\right) I_L(\hat{\eta}_L)(\hat{\eta}_L - \eta_L) \\ &= X'\hat{V}^{-1}(\hat{\eta} - \eta), \end{aligned} \tag{16}$$

TABLE 1 *A 2 × 2 Table with Supplemental Margins for both variables*

Set	Y_1	Y_2	Count
H	1	1	100
	1	2	50
	2	1	75
	2	2	75
K	1		30
	2		60
L		1	28
		2	60

where $\eta = (\eta_H', \eta_K', \eta_L')'$ and $\hat{\eta} = (\hat{\eta}_H', \hat{\eta}_K', \hat{\eta}_L')'$. Therefore, inserting (14) and (16) into (10), we have

$$\theta^{(k+1)} = \theta^{(k)} + \left[X'\hat{V}^{-1}X\right]^{-1} X'\hat{V}^{-1}\left[\hat{\eta} - \eta\left(\theta^{(k)}\right)\right], \tag{17}$$

which is equivalent to the expression in (8).

4. A Numerical Example

For a numerical example, we consider the data set originally presented by Little (1982) and also discussed in Little and Rubin (2002). Table 1 gives the data for a 2 × 2 table with supplemental margins for both the classifying variables. According to Little (1982), the final probabilities of classification obtained from EM algorithm are

$$\hat{\pi}_{11} = 0.28, \quad \hat{\pi}_{12} = 0.17, \quad \hat{\pi}_{21} = 0.24, \quad \hat{\pi}_{22} = 0.31, \tag{18}$$

where $\pi_{ij} = Pr\left(Y_1 = i, Y_2 = j\right), i, j = 1, 2$.

For the orthogonal parametrization, we use

$$\boldsymbol{\eta}_H = \left(\pi_{1|1}, \pi_{1|2}, \pi_{+1}\right)'$$

where $\pi_{1|1} = Pr\left(Y_1 = 1 \mid Y_2 = 1\right)$, $\pi_{1|2} = Pr\left(Y_1 = 1 \mid Y_2 = 2\right)$, $\pi_{+1} = Pr\left(Y_2 = 1\right)$. We also set $\boldsymbol{\theta} = \boldsymbol{\eta}_H$. Note that the validity of the proposed method does not depend on the choice of the parametrization. A suitable parametrization will make the computation of the information matrix simple.

From the data in Table 1, the five observations for three parameters are

$$\begin{aligned} \hat{\boldsymbol{\eta}} &= \left(\hat{\pi}_{1|1,H}, \hat{\pi}_{1|2,H}, \hat{\pi}_{+1,H}, \hat{\pi}_{1+,K}, \hat{\pi}_{+1,L}\right)' \\ &= \left(100/175, 50/125, 175/300, 30/90, 28/88\right)' \end{aligned}$$

with the expectations

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \left(\pi_{1|1}, \pi_{1|2}, \pi_{+1}, \pi_{1|1}\pi_{+1} + \pi_{1|2} - \pi_{1|2}\pi_{+1}, \pi_{+1}\right)'$$

and the variance-covariance matrix

$$\mathbf{V} = diag\left\{\frac{\pi_{1|1}\left(1-\pi_{1|1}\right)}{n_H}, \frac{\pi_{1|2}\left(1-\pi_{1|2}\right)}{n_H}, \frac{\pi_{+1}\left(1-\pi_{+1}\right)}{n_H}, \frac{\pi_{1+}\left(1-\pi_{1+}\right)}{n_K}, \frac{\pi_{+1}\left(1-\pi_{+1}\right)}{n_L}\right\}.$$

The Gauss-Newton method as in (6) can be used to solve the nonlinear model of three parameters, where the initial estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}_S} = (100/175, 50/125, 203/388)'$ and the X matrix is

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & \pi_{+1} & 0 \\ 0 & 1 & 0 & 1-\pi_{+1} & 0 \\ 0 & 0 & 1 & \pi_{1|1}-\pi_{1|2} & 1 \end{pmatrix}'. \tag{19}$$

The resulting one-step estimates are

$$\hat{\pi}_{11} = 0.29, \quad \hat{\pi}_{12} = 0.18, \quad \hat{\pi}_{21} = 0.23, \quad \hat{\pi}_{22} = 0.30,$$

which is close to the final results in (18) obtained from the EM algorithm.

References

ANDERSON, T.W. (1957). "Maximum likelihood estimates for the multivariate normal distribution when some observation are missing", *Journal of the American Statistical Association*, **52**, 200–203.

DEMPSTER, A.P., LAIRD, N.M., AND RUBIN, D.B. (1977). "Maximum likelihood from incomplete data via the EM algorithm (with discussion)", *Journal of Royal Statistical Society, Series B*, **39**, 1–38.

FULLER, W.A. (1996). *Introduction to statistical times series,* Wiley, New York.

LEHMANN, E.L. (1983). *Theory of Point Estimation*, Wiley, New York.

LITTLE, R.J.L. (1982). "Models for nonresponse in sample surveys", *Journal of the American Statistical Association*, **77**, 237–250.

LITTLE, R.J.L. AND RUBIN, D.B. (2002). *Statistical Analysis with missing data*, Wiley, New York.

RUBIN, D.B. (1974). "Characterizing the estimation of parameters in incomplete data problems", *Journal of the American Statistical Association*, **69**, 467–474.

RUBIN, D.B. (1976). "Inference and missing data" *Biometrika,* **63**, 581–590.